Psychological Monographs

General and Applied

No. 439
1957

Newman, Howell, and Harris

# Forced Choice and Other Methods for Evaluating Professional Health Personnel

By

Sidney H. Newman, Margaret A. Howell

*United States Public Health Service*

and Frank J. Harris

*Operations Research Office, the Johns Hopkins University*

Vol. 71
No. 10

Price $1.00

Edited by Herbert S. Conrad

Published by The American Psychological Association, Inc.

# Psychological Monographs: General and Applied

Combining the *Applied Psychology Monographs* and the *Archives of Psychology* with the *Psychological Monographs*

HERBERT S. CONRAD, Editor

*Department of Health, Education, and Welfare*
*Office of Education*
*Washington 25, D.C.*



MANUSCRIPTS should be sent to the Editor.

Because of lack of space, the *Psychological Monographs* can print only the original or advanced contribution of the author. *Background and bibliographic materials must, in general, be totally excluded* or kept to an irreducible minimum. Statistical tables should be used to present only the most important of the statistical data or evidence.

The first page of the manuscript should contain the title of the paper, the author's name, and his institutional connection (or his city of residence). Acknowledgments should be kept brief, and appear as a *footnote* on the first page. No table of contents need be included. For other directions or suggestions on the preparation of manuscripts, see: CONRAD, H. S. Preparation of manuscripts for publication as monographs. *J. Psychol.*, 1948, 26, 437-459.

CORRESPONDENCE CONCERNING BUSINESS MATTERS (such as subscriptions and sales, change of address, etc.) should be addressed to the American Psychological Association, Inc., 1333 Sixteenth St. N.W., Washington 6, D.C. Address changes must arrive by the 10th of the month to take effect the following month. Undelivered copies resulting from address changes will not be replaced; subscribers should notify the post office that they will guarantee third-class forwarding postage.

Vol. 71, No. 10 Whole No. 439, 1957

Psychological Monographs: General and Applied

# Forced Choice and Other Methods for Evaluating Professional Health Personnel

SIDNEY H. NEWMAN, MARGARET A. HOWELL
*United States Public Health Service*

AND FRANK J. HARRIS
*Operations Research Office, The Johns Hopkins University*[1]

## INTRODUCTION

THIS study was undertaken to compare the forced choice technique with other methods for evaluating the performance of commissioned professional health personnel in the United States Public Health Service. It is a part of the officer selection and evaluation program described by Newman (7).

The Public Health Service, the major Federal organization responsible for the health of the nation, employs approximately 16,000 Civil Service personnel and 3,000 commissioned officers. The commissioned officer component of the Service is composed of carefully selected personnel in various scientific specialties and in the health professions of medicine, dentistry, nursing, sanitary engineering, pharmacy, veterinary medicine, dietetics, and physical therapy. These officers hold clinical, research, public health, and administrative positions in Service hospitals, outpatient clinics, research centers, regional offices, other governmental agencies, and foreign countries.

The performance evaluation of commissioned officers is accomplished through periodic efficiency reporting. In an effort to improve the Service's performance-rating system, an investigation of methods for evaluating job performance was undertaken in 1949. A review of the literature in the performance-rating field revealed that no data had been reported then, or at present, on the kinds of highly trained scientific and professional personnel employed in the varied and specialized work areas in the Service.

A consideration of various performance evaluation methods led to the conclusion that the forced choice technique developed by the Department of the Army appeared most promising for use in the Public Health Service setting. While investigations of the forced choice technique have been based on populations which are quite different from Public Health Service professional personnel, reports by Sisson (10), Witsell (12), and the Adjutant General's Office, Department of the Army (13, 14) seemed to indicate the usefulness of the technique in a commissioned personnel system.

An Experimental Efficiency Report, incorporating forced choice items and other evaluation materials not in use in the Service in 1949, was designed and distributed to the supervisors of active-duty officers. The effectiveness of the forced choice technique is here compared with that of the more conventional evaluation methods included in the Experimental Report. It is anticipated that the results of this study will contribute to the literature on performance evaluation and, in particular, to that of forced choice methodology. The findings may also be of special interest to local or state health departments, hospitals, research institutions, or other organizations employing professional personnel engaged in public health work, medical care, or research relevant to the problems of

[1] Formerly with the United States Public Health Service.

health and disease. In addition, this study calls attention to some of the problems implicit in the evaluation of relatively small numbers of employees engaged in a variety of specialized professional activities.

## Problems

An investigation of performance-evaluation methods was undertaken to answer the following questions:

1. Is the forced choice technique an effective method for measuring the performance of professional health personnel in the Public Health Service?

2. How does the forced choice technique compare in validity and reliability with more conventional methods of performance evaluation?

3. How do factors such as the characteristic evaluated in criterion ratings, the administrative level of the supervisor completing efficiency reports, and the grade of the ratee affect the validity of efficiency reporting?

4. What combination of efficiency-reporting methods optimally predicts the performance of personnel in the various professional and occupational fields of the Service?

In addition to the major problems of the study, it was possible to compare the validity of the Experimental Report with that of the Officer's Progress Report (the efficiency report in use in the Service in 1949) which had been completed under operational rather than experimental conditions.

## Materials

### *Criteria*

The criteria of performance within the Public Health Service consisted of 20-point graphic rating scales used for the evaluation of each of the following factors: Work Performance, Administrative Ability, Personality (Personal Qualifications), and Over-all Value to the Service. Instructions for each scale requested raters to compare the ratee with a typical group of personnel having similar duties and responsibilities. A rating of 1 was used to designate the least effective ratees, and a rating of 20 was used to designate the most effective.

### *Experimental Efficiency Report*

This Report was divided into the following four sections, samples of which may be seen in the Appendix.

*Section I—Forced Choice.* This part consisted of 50 tetrads adapted from items developed by the Department of the Army (10).[2] Each tetrad was composed of four words or phrases descriptive of job performance or personal qualifications from which a supervisor was to select (*a*) the one most descriptive and (*b*) the one least descriptive of the individual he was rating. A preliminary investigation had shown that the Army tetrads adapted for use with Public Health Service personnel produced a promising number of scorable alternatives (9).

*Section II—Job Proficiency.* This was a list of ten major work areas in the Public Health Service from which a supervisor was to indicate the ratee's primary job function. The supervisor was then requested to rate, on a ten-point scale, the quality of the ratee's performance in this function.

*Section III—Personal Qualifications.* This section consisted of ten-point rating scales for the evaluation of eight personality characteristics such as reaction to criticism, freedom from bias and emo-

[2] Appreciation is expressed to the Personnel Research Branch, The Adjutant General's Office, Department of the Army, for making these materials available.

tional upset, ability to work with others, ability to act on own responsibility, and diligence and persistence in performing necessary work.

*Section IV—Check List.* This part consisted of 22 statements which were to be marked as applying or not applying to the ratee. The statements concerned professional knowledge, interest in work, planning and organizing ability, leadership, versatility, and other characteristics related to work performance. In the development of the Check List, some 500 statements had been extracted verbatim from "Remarks" sections of the Officer's Progress Report, placed in 12 logical categories, and sorted according to Thurstone's variation of the method of equal-appearing intervals (8, 11). The 22 statements comprising the final check list were those which showed the least variability (smallest semi-interquartile range) and which were deemed most relevant to performance.

### *Officer's Progress Report*

Samples of parts of the Progress Report used in 1949 to obtain periodic efficiency ratings on commissioned personnel are shown in the Appendix. The Progress Report contains two types of evaluations:

*Rating Scales.* This section consists of 11 five-point rating scales for evaluating such factors as judgment, general professional knowledge, proficiency in assigned duties, industry, tact, initiative, and dependability. The scales are scored by a point system, odd number values of one through nine being assigned to the five points on each scale; values from all scales are averaged to obtain a total score.

*Narrative Comments.* Several questions in the Progress Report also elicit narrative comments concerning a ratee's performance. A method for scoring these comments was developed (8). Its use involves assigning to a comment in a Progress Report the scale value of a matching comment in a scoring manual. The total score for the Narrative Comments is the average of the scale values of all comments in a Report.

Total raw scores from the Narrative Comments and the Rating Scales are separately converted to standard scores on the basis of norms established for each officer grade and profession.[3]

## Collection of Data

Criterion ratings were obtained during 1949 from 45 of the 54 Public Health Service installations in the United States, including 14 hospitals, 10 regional offices, 8 divisions, 8 laboratories of the National Institutes of Health, and 5 other installations such as outpatient clinics. Nine stations were excluded from the study because of practical considerations and the small numbers of possible ratees at most of these stations.

Ratings were obtained in a systematic manner by a staff representative who explained the method of rating and administered the forms. Officers who worked together, regardless of profession or grade, met in groups and rated each other. For each of the rating factors, which were randomly alternated, each officer was provided with a roster of all officers at his station. He was then asked to rate each officer, excluding himself and officers he did not know, by placing the ratee at one of the points on the twenty-point scale. Ratings were performed anonymously with the assurance that they were to be used for research purposes only. The lower grades of officers (the equivalents of the Navy ensign through full lieutenant) were rated on

[3] Public Health Service officer grades and their Navy equivalents are given below:

| *Public Health Service* | *Navy* |
|---|---|
| Junior assistant | Ensign |
| Assistant | Lieutenant (j.g.) |
| Senior assistant | Lieutenant |
| Full | Lieutenant commander |
| Senior | Commander |
| Director | Captain |

one scale and the higher grades on another in an effort to reduce irrelevant grade-associated factors which might affect the ratings.

One month after the collection of criterion data, copies of the Experimental Efficiency Report, directions for completing, and a schedule for designating supervisors to mark the Experimental Reports were mailed to the officers in charge of the installations which had been visited for purposes of collecting criterion ratings. Two independent Reports were requested on each ratee, one from his immediate officer supervisor and another from either the officer in charge or his representative.

The Officer's Progress Report is requested annually for all officers at the Full grade (Navy lieutenant commander) and above, and semiannually for all officers below this grade. For purposes of the present study, one Progress Report was selected, where possible, for each officer on whom an Experimental Report had been completed. The Progress Reports selected were those completed within six months before or after the Experimental Reports. In most instances, this time control resulted in a matching of the two Reports on ratee's Service profession, grade, corps, and station. Progress Reports which did not match the Experimental Reports on these factors were eliminated from the study.

## Development of Experimental Report Scoring Keys

### *Designation of Occupational Groups for Item Analysis*

The diversity of professions and job functions within the Public Health Service necessitated the designation of separate occupational groups for which Experimental Report scoring keys could be developed. An early analysis of the Experimental Report showed that items scorable for various ratee professions did not appreciably overlap (9). In view of these considerations, the following factors were controlled in designating groups for item-analysis purposes.

*Criteria.* Intercorrelations among scores on the four criteria (not shown in tabular form) revealed that Work Performance and Personality produced the lowest intercorrelations, ranging from .50 to .72; all other correlations were higher, ranging from .74 to .92. The decision was made to use only the two more independent criteria, Work Performance and Personality, for purposes of item analyzing the Experimental Report.

*Station.* This factor was controlled by type of station. Stations were classified into three groups according to their major functions: (*a*) medical care, furnished in installations such as hospitals and outpatient clinics; (*b*) public health work, carried on in regional offices and in such divisions as those of the Bureau of State Services; and (*c*) research, performed in installations such as the National Institutes of Health.

*Profession of rater.* Among medical care personnel (in hospitals and outpatient clinics), where the number of criterion ratings permitted, profession of rater was controlled by using only those ratings performed by members of the ratee's profession. For nurses working in medical care, two groups of raters were used: physicians and nurses. In the public health and the research groups, established by the station control, ratings by all professions of raters were used because of the small numbers of personnel in any one profession. Further, public health and research personnel are usually given efficiency ratings by supervisors working in the same functional area, but not in the same profession. Medical care personnel, however, not only receive efficiency ratings from supervisors in the medical care field, but frequently from those in the particular profession of the ratee.

*Profession of ratee.* For purposes of item analysis, this variable was controlled among medical care personnel by the establishment of separate groups according to the three major professions represented—medicine, dentistry, and nurs-

TABLE 1

MEANS AND STANDARD DEVIATIONS OF CRITERION SCORES IN EACH OCCUPATIONAL GROUP DESIGNATED FOR ITEM ANALYSIS

| Occupational Group | Work Performance Criterion | | | Personality Criterion | | |
|---|---|---|---|---|---|---|
| | No. Ratees | M[b] | *SD* | No. Ratees[c] | M | *SD* |
| Physicians (1)[a] | 158 | 12.42 | 2.92 | 159 | 11.77 | 2.88 |
| Physicians (2) | 161 | 12.54 | 2.72 | 161 | 11.84 | 2.60 |
| Public health personnel (1) | 90 | 13.23 | 2.43 | 99 | 12.75 | 2.35 |
| Public health personnel (2) | 88 | 13.60 | 2.34 | 100 | 13.23 | 2.25 |
| Research personnel (1) | 66 | 13.88 | 2.52 | 69 | 12.97 | 2.33 |
| Research personnel (2) | 65 | 13.79 | 2.63 | 67 | 12.50 | 3.08 |
| Nurses rated by nurses (1) | 56 | 13.28 | 2.80 | 56 | 12.66 | 3.04 |
| Nurses rated by nurses (2) | 55 | 13.42 | 2.64 | 55 | 12.84 | 2.80 |
| Nurses rated by physicians | 92 | 12.44 | 2.17 | 91 | 11.51 | 2.22 |
| Dentists | 60 | 13.29 | 1.98 | 62 | 12.90 | 2.10 |

[a] (1) = sample 1; (2) = sample 2.

[b] Mean scores from matched samples did not differ significantly at the .05 level or below.

[c] In some groups more ratees were evaluated on the Personality than on the Work Performance criterion, perhaps because raters felt they had greater opportunity to observe personal qualifications than job performance.

ing. The groups of public health and research personnel were not broken down by profession of ratee for the same reasons as those specified in the discussion of profession of rater.

*Grade of ratee.* Grade was controlled by the proportionate representation of each ratee grade in high, middle, and low criterion groups identified within each of the separate item-analysis groups. The criterion groups were the upper 27 per cent, the middle 46 per cent, and the lower 27 per cent of ratees, determined by the average rating they received from the appropriate group of raters on the separate criteria of Work Performance and Personality.

Within the occupational field of medical care, then, four groups were established for purposes of item analysis: (*a*) physicians rated by physicians; (*b*) nurses rated by nurses; (*c*) nurses rated by physicians; and (*d*) dentists rated by dentists and physicians. Two other item-analysis groups based on occupational fields were also established: (*a*) public health personnel rated by public health personnel; and (*b*) research personnel rated by research personnel.

A criterion score, the average of five or more ratings, was computed for each ratee on each criterion. A minimum of five ratings was required to obtain as highly reliable scores as possible without excluding from the study a large number of the ratees. Where numbers permitted, the groups designated for item analysis were split into matched samples to provide for cross validation. All groups except the dentists and the nurses rated by physicians contained enough ratees to furnish split samples.

The number in each item-analysis group, as well as the mean and the standard deviation of the criterion scores on each rating factor, is shown in Table 1. From Table 1, it is important to note that none of the mean differences in matched groups is significant (five per cent level or below).

### *Item Analysis*

For purposes of item analysis, one Experimental Report was selected for each ratee. The one Report selected, termed the "primary" Report, was gen-

erally completed by a supervisor whose administrative level corresponded to that of a branch chief in a division or a clinical director in a hospital. A "secondary" Report was available on most ratees; this Report was generally completed by a supervisor, such as the officer in charge, who was at a higher organizational level than the supervisor completing the primary Report. Secondary Reports, although not used in item analysis, were scored to provide additional validation data.

*Section I—Forced Choice.* In a previous study, it was found that of four methods for item analyzing forced choice tetrads, the critical-ratio technique appeared the most useful in that it was relatively easy to apply, gave readily interpretable results, and yielded item weights which were in close agreement with weights derived from the other methods studied (9). In the present work, critical ratios were used to item analyze the Forced Choice section against the separate criteria of Work Performance and Personality. Within each item-analysis group, the significance of the difference was tested between the percentage of the high and the percentage of the low criterion groups rated on each alternative. In the medical groups, the samples were sufficiently large that it was possible to use a critical ratio of 1.96 as the standard for scoring. In the remaining smaller occupational groups, an alternative was deemed scorable if the critical ratio were 1.50 or greater. Unitary positive weights, indicating that a significantly higher percentage of a high criterion group than of a low criterion group had been rated on a given alternative, and unitary negative weights, indicating the reverse, were assigned the scorable alternatives in the tetrads. Alternatives which were nondiscriminating received zero weights.

From the scorable tetrads for each item-analysis group, approximately the best 20 were selected to constitute scoring keys. In addition, for each of the item-analysis groups which had been split, a combined sample scoring key was developed. This key was composed of the best 20 tetrads selected from those in which only alternatives having identical scoring weights in the matched samples had been retained. For the group of nurses rated by nurses, only a combined sample key was developed since the matched samples did not individually yield enough scorable items for separate keys.

A total score on the Forced Choice section was obtained by summing all positively weighted alternatives. A previous study indicated that positive weights scoring yielded as valid results as positive plus negative weights scoring on three lengths of keys, one of which was a twenty-tetrad key. The validity of this length of key also compared favorably with that of the other two key lengths studied (5).

*Sections II and III—Job Proficiency and Personal Qualifications.* Since these two sections consisted of ten-point rating scales, the same methods of item analysis were used for both. First, the discriminatory capacity of the scales was checked by testing the significance of the difference in the mean scale values of high and of low criterion groups. Of the total number of critical ratios computed on the Personal Qualifications scales, considering all item-analysis groups, 70.6 per cent (113 out of 160) were significant at the .05 level or below. On the single Job Proficiency scale, seven of the item analysis groups yielded significant differences (.05 level or below) between upper and lower 27 per cent groups on one or both criteria.

The raw scores of all ratees in each

item-analysis group were used to develop stanine scores in one-half sigma units, with the mean scale value equalling a stanine of five. The stanine scoring resulting for matched samples was so similar that the split samples were recombined to lend greater stability to the stanine scales. Cross validation of the Job Proficiency and Personal Qualifications sections appeared unnecessary in view of the consistency in scoring from one matched sample to another.

The total score on the Personal Qualifications section was the average of the stanines from the eight rating scales. The score on Job Proficiency was the stanine value of the rating given a ratee in his primary job function.

While it would have been desirable to treat separately each of the 10 functions listed in the Job Proficiency section, this was not possible because of the small number of ratees performing each function. All analyses on this section were made without regard to the type of work involved. Specific job functions were not completely masked, however, in view of the method used in establishing item-analysis groups. For example, of the 179 ratees in the public health groups who were rated on the Work Performance criterion, 138 were given ratings on the Experimental Efficiency Report in the primary function described as "operation in a technical or specialized Public Health program."

*Section IV—Check List.* Critical ratios were computed to test the significance of the difference in the percentages of high and of low criterion groups marked on each alternative. Items which discriminated between the two groups at the .05 level or below were deemed scorable. Considering both criteria and all item-analysis groups, 42.3 per cent (186 out of 440) of the ratios computed reached this level of significance; the scorable items, however, were not evenly distributed among the various item-analysis groups.

Considering the separate item-analysis groups, the number of scorable items was such that it was feasible to develop scoring keys against the Work Performance criterion in only the medical, public health, and research groups, and against the Personality criterion in only the medical groups. Since these were split groups, the decision was made to include in the scoring keys only those items that reached the required level of significance in both of the matched samples rather than to develop scoring keys for the separate samples.

The total score on this section was the sum of all positively weighted items (those characteristic of a high criterion group).

## Results and Interpretation

### *Comparison of Forced Choice Scoring Keys*

The Forced Choice section of the Experimental Reports from each of the split sample item-analysis groups was scored by three keys: (*a*) self scoring, developed from item analysis of the sample being scored; (*b*) cross scoring, developed from item analysis of the matched sample and used for cross validation; and (*c*) combined sample scoring, based on alternatives that gave the same scoring weights in both samples.

Validity coefficients based on each type of Forced Choice key are presented in Table 2 for the matched sample groups. Data on the three scoring keys are given for secondary Reports as well as for the primary Reports used in item analysis. From the validity coefficients in Table 2, it may first be noted that the coefficients were highly similar from one matched sample to another. In comparisons of matched sample validities, only

TABLE 2

FORCED CHOICE VALIDITY COEFFICIENTS FOR MATCHED SAMPLE GROUPS[a]

| Occupational Group | Work Performance Criterion | | | | Personality Criterion | | | |
|---|---|---|---|---|---|---|---|---|
| | No. Ratees | Self Scoring | Cross Scoring | Combined Scoring | No. Ratees | Self Scoring | Cross Scoring | Combined Scoring |
| | Primary Reports | | | | | | | |
| Physicians (1)[e] | 158 | .67 | .65 | .66 | 159 | .71[d] | .61 | .69[e] |
| Physicians (2) | 161 | .58 | .55 | .56 | 161 | .55 | .56 | .54 |
| Public health personnel (1) | 90 | .62[df] | .50 | .57[e] | 99 | .53 | .56 | .58 |
| Public health personnel (2) | 88 | .61 | .55 | .59 | 100 | .55 | .56 | .56 |
| Research personnel (1) | 66 | .67 | .63 | .69[e] | 69 | .65 | .55 | .62 |
| Research personnel (2) | 65 | .67[d] | .57 | .62 | 67 | .60 | .56 | .64 |
| Median $r$ | | .65 | .56 | .61 | | .58 | .56 | .60 |
| | Secondary Reports | | | | | | | |
| Physicians (1) | 120 | .67 | .66 | .67 | 120 | .70 | .69 | .70 |
| Physicians (2) | 123 | .70 | .67 | .68 | 123 | .63 | .66 | .66 |
| Public health personnel (1) | 64 | .38[f] | .34 | .30[b] | 71 | .47 | .49 | .52 |
| Public health personnel (2) | 63 | .47 | .44 | .45 | 68 | .42 | .50 | .47 |
| Research personnel (1) | 51 | .40 | .31[b] | .37 | 53 | .64 | .55 | .59 |
| Research personnel (2) | 56 | .53 | .52 | .57 | 58 | .50 | .60 | .58 |
| Median $r$ | | .50 | .48 | .51 | | .57 | .58 | .59 |

[a] All $r$s not marked are significantly different from zero at the .01 level or below.
[b] $r$ is significantly different from zero at the .05 level.
[c] (1) = sample 1; (2) = sample 2.
[d] $r$ on self scoring is significantly higher at the .01 level than $r$ on cross scoring.
[e] $r$ on combined scoring is significantly higher at the .05 level than $r$ on cross scoring.
[f] $r$ on self scoring is significantly higher at the .05 level than $r$ on combined scoring.

two differences significant at the .05 level or below occurred. These were on primary Reports, Personality criterion, in the comparison of physicians (1) and (2)[4] on the self ($r$s = .71 vs. .55) and the combined scoring keys ($r$s = .69 vs. .54).[5]

Table 2 also shows that the self and the combined keys, both of which represent the use of scoring keys with item-analysis groups, tended to produce higher validity coefficients than did scoring keys used with independent samples established for cross validation. This trend is apparent both from the median correlations and from the tests of differences in validity from one type of scoring to another.[6] It should be noted that the differences in validity by type of scoring appeared to decrease when scoring keys were applied to Reports (secondary) independent of those used in item analysis. Only one significant difference in validity by type of scoring occurred on the secondary Reports.

Correlations (not shown in the table) among the three types of scoring keys were high. Considering coefficients based on both primary and secondary Reports,

[4] Here, (1) = sample 1; (2) = sample 2.

[5] For purposes of testing the significance of the difference in $r$s, $r$s were transformed to $z$s. Tests of differences reported in this paper involved independent samples and the same sample with one array in common (4, p. 124, formulas 45, 47, and 49).

[6] Median correlation coefficients have been presented in tables merely to aid the reader in observing trends in the data. They are not intended to be precise summary statistics since an assumption that the various officer groups were samples drawn from a common population is not warranted.

TABLE 3

VALIDITY COEFFICIENTS FOR ALL SECTIONS OF THE EXPERIMENTAL EFFICIENCY REPORT[a]

| Occupational Group | Work Performance Criterion | | | | | Personality Criterion | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | No. Ratees | FC | JP | PQ | CL | No. Ratees | FC | JP | PQ | CL |
| | Primary Reports | | | | | | | | | |
| Physicians | 319 | .61 | .62 | .58 | .54 | 320 | .62 | .40 | .50 | .46 |
| Public health personnel | 178 | .58 | .49 | .49 | .54 | 199 | .57 | .20 | .36 | |
| Research personnel | 131 | .65 | .44 | .39 | .59 | 136 | .63 | .18[b] | .29 | |
| Nurses rated by nurses | 111 | .42 | .31 | .39 | | 111 | .43 | .22[b] | .35 | |
| Nurses rated by physicians | 92 | .44 | .28 | .43 | | 91 | .37 | .20[c] | .36 | |
| Dentists | 60 | .53 | .49 | .55 | .50 | 62 | .35 | .41 | .40 | .33 |
| Median $r$ | | .56 | .47 | .46 | .54 | | .50 | .26 | .36 | .40 |
| | Secondary Reports | | | | | | | | | |
| Physicians | 243 | .67 | .58 | .60 | .60 | 243 | .68 | .52 | .58 | .56 |
| Public health personnel | 127 | .37 | .37 | .37 | .49 | 139 | .50 | .30 | .33 | |
| Research personnel | 107 | .48 | .47 | .41 | .49 | 111 | .58 | .32 | .39 | |
| Nurses rated by nurses | 91 | .40 | .24[b] | .35 | | 91 | .53 | .07[c] | .22[b] | |
| Nurses rated by physicians | 73 | .52 | .22[c] | .34 | | 72 | .45 | .10[c] | .27[b] | |
| Dentists | 42 | .66 | .55 | .48 | .64 | 43 | .50 | .50 | .35[b] | .49 |
| Median $r$ | | .50 | .42 | .39 | .55 | | .52 | .31 | .34 | .53 |

[a] All $r$s not marked are significantly different from zero at the .01 level or below.
[b] $r$ is significantly different from zero at the .05 level.
[c] $r$ does not reach the .05 level of significance.

the median correlation between the self and the cross keys was .92, between the self and the combined, .96, and between the cross and the combined, .96. Both the self and the cross scoring keys correlated highly with the combined key since the latter was composed of the tetrad alternatives scored in both samples.

It was to be expected that the self and the combined keys would yield the higher validity coefficients since they were used to score item-analysis Reports. However, when the self keys were applied as cross keys to Experimental Reports independent of those used in item analysis, the validities exhibited surprisingly little decrease in size. Out of 24 possible comparisons of the self and the cross keys, only three (12.5 per cent) were significant at the .05 level or below. Although cross-validation data were not available for the combined keys, it seems likely that in successive samples they would have greater stability than keys derived from item analysis of Reports completed on a single sample. For this reason, subsequent discussion of the Forced Choice section of the Experimental Report will be based only on data from the combined sample scoring keys. In order to increase the reliability of the statistics based on these keys, matched samples have been recombined.

### *Validity of the Experimental Report*

Table 3 presents the validity coefficients based on the Forced Choice (FC),

Job Proficiency (JP), Personal Qualifications (PQ), and Check List (CL) sections of the Experimental Report.[7] To facilitate interpretation of the correlations, the factors which may be influencing variations in the data will be individually considered.

*Report sections.* From the median correlations in Table 3, it appears that the validity coefficients for the Forced Choice and Check List sections were higher than those obtained for the Job Proficiency and the Personal Qualifications sections. More specific comparisons of those Experimental Report sections which showed significant differences in validity may be seen in Table 4.

Out of the 108 possible comparisons of Experimental Report sections, 36 (33.3 per cent) were significant at the .05 level or below. In 27 of the 36 significantly different pairs of coefficients, the Forced Choice section exhibited higher validities. In only one instance was the validity of the Forced Choice section significantly lower than that of another section. The Job Proficiency and Personal Qualifications scales produced the lowest coefficients; in 15 instances, each of these sections yielded a validity coefficient which was significantly lower than that of another section. The Job Proficiency scale in only three instances and the Personal Qualifications section in only two instances produced validities which were significantly higher than those of other Report sections. The Check List in four comparisons yielded a significantly higher coefficient than another section, and in five comparisons, a significantly lower coefficient. In four of the five instances in which the Check List produced a lower coefficient, it was compared with the Forced Choice section of the Report.

In general, the Forced Choice section gave the highest validity coefficients. Comparisons of validities on the remaining three sections of the Experimental Report yielded relatively few significant differences although the Check List, where available, tended to produce somewhat higher validities than did the Job Proficiency and Personal Qualifications sections.

TABLE 4

COMPARISONS OF SECTIONS OF THE EXPERIMENTAL REPORT IN WHICH VALIDITY COEFFICIENTS DIFFERED SIGNIFICANTLY

| Occupational Group | Primary Reports | Secondary Reports |
|---|---|---|
| | Work Performance Criterion | |
| | Report Sections Compared[a] | Report Sections Compared[a] |
| Physicians | FC vs. CL[c]<br>JP vs. CL[c] | FC vs. JP[b]<br>FC vs. PQ[c]<br>FC vs. CL[b] |
| Public health personnel | | CL vs. FC[c] |
| Research personnel | FC vs. JP[b]<br>FC vs. PQ[b]<br>CL vs. JP[c]<br>CL vs. PQ[b] | |
| Nurses rated by physicians | | FC vs. JP[b]<br>FC vs. PQ[c] |
| Dentists | | FC vs. PQ[c]<br>CL vs. PQ[c] |
| | Personality Criterion | |
| Physicians | FC vs. JP[b]<br>FC vs. PQ[b]<br>FC vs. CL[b]<br>PQ vs. JP[b] | FC vs. JP[b]<br>FC vs. PQ[b]<br>FC vs. CL[b]<br>JP vs. PQ[c] |
| Public health personnel | FC vs. JP[b]<br>FC vs. PQ[b] | FC vs. JP[c]<br>FC vs. PQ[c] |
| Research personnel | FC vs. JP[b]<br>FC vs. PQ[b]<br>PQ vs. JP[c] | FC vs. JP[c]<br>FC vs. PQ[c] |
| Nurses rated by nurses | FC vs. JP[c] | FC vs. JP[b]<br>FC vs. PQ[b] |
| Nurses rated by physicians | | FC vs. JP[c] |
| Dentists | | JP vs. PQ[c] |

[a] The rating section on which the higher validity was obtained is listed first.

[b] Validity coefficients for the sections compared differ significantly at the .01 level or below.

[c] Validity coefficients for the sections compared differ significantly at the .05 level.

[7] Reports in the dental group were scored by the key developed on physicians. An exploratory validation study showed as high validities for dentists as for physicians when the medical key was used to score reports in both groups.

TABLE 5

COMPARISONS OF OCCUPATIONAL GROUPS IN WHICH VALIDITY COEFFICIENTS DIFFERED SIGNIFICANTLY

| Work Performance Criterion | | | | Personality Criterion | | | |
|---|---|---|---|---|---|---|---|
| Groups Compared[a] | | | Report Section | Groups Compared[a] | | | Report Section |
| **Primary Reports** | | | | | | | |
| Physicians | vs. | Nurses rated by nurses | FC[c] | Physicians | vs. | Nurses rated by nurses | FC[c] |
| | | Nurses rated by phys. | FC[c] | | | Nurses rated by phys. | FC[b] |
| | | P. h. personnel | JP[c] | | | Dentists | FC[c] |
| | | Res. personnel | JP[c] | | | Res. personnel | JP[c] |
| | | Nurses rated by nurses | JP[b] | | | Res. personnel | PQ[c] |
| | | Nurses rated by phys. | JP[b] | | | | |
| | | Res. personnel | PQ[c] | P. h. personnel | vs. | Nurses rated by phys. | FC[c] |
| | | Nurses rated by nurses | PQ[c] | | | Dentists | FC[c] |
| Res. personnel | vs. | Nurses rated by nurses | FC[c] | Res. personnel | vs. | Nurses rated by nurses | FC[c] |
| | | Nurses rated by phys. | FC[c] | | | Nurses rated by phys. | FC[c] |
| | | | | | | Dentists | FC[c] |
| **Secondary Reports** | | | | | | | |
| Physicians | vs. | P. h. personnel | FC[b] | Physicians | vs. | P. h. personnel | FC[b] |
| | | Res. personnel | FC[c] | | | Nurses rated by phys. | FC[c] |
| | | Nurses rated by nurses | FC[b] | | | P. h. personnel | JP[c] |
| | | P. h. personnel | JP[c] | | | Res. personnel | JP[c] |
| | | Nurses rated by nurses | JP[b] | | | Nurses rated by nurses | JP[b] |
| | | Nurses rated by phys. | JP[b] | | | Nurses rated by phys. | JP[b] |
| | | P. h. personnel | PQ[b] | | | P. h. personnel | PQ[b] |
| | | Res. personnel | PQ[c] | | | Res. personnel | PQ[c] |
| | | Nurses rated by nurses | PQ[b] | | | Nurses rated by nurses | PQ[b] |
| | | Nurses rated by phys. | PQ[c] | | | Nurses rated by phys. | PQ[b] |
| Dentists | vs. | P. h. personnel | FC[c] | Dentists | vs. | Nurses rated by nurses | JP[c] |
| | | Nurses rated by phys. | JP[c] | | | Nurses rated by phys. | JP[c] |

[a] The group in which the higher validity was obtained is listed first.
[b] Validity coefficients for the groups compared differ significantly at the .01 level or below.
[c] Validity coefficients for the groups compared differ significantly at the .05 level.

*Occupational group.* From Table 3, the significance of the difference was also tested between the validity coefficients obtained from one occupational group to another. The comparisons yielding significant differences are summarized in Table 5. It should be mentioned that tests of differences were only made between independent occupational groups. The two nursing groups, which overlapped in membership, were not compared since the amount of computational work involved did not seem warranted by the relatively small differences in validity that occurred in most instances between the two groups.

Out of the 182 group comparisons made, Table 5 shows that 44 (24.2 per cent) yielded differences significant at the .05 level or below. It is rather striking that in 33 of the significant comparisons, the medical group produced the higher validity coefficient, while in no instance did it produce a significantly lower one. In 25 of the significant comparisons, the lower coefficient occurred in one of the two nursing groups. Coefficients in the public health, research, and dental groups tended to be of about the same magnitude, differing in some instances from the two extreme groups of physicians and nurses. No significant differences were found between the public health and the research groups; only one significant difference occurred in the comparisons of dentists and research personnel, and two in the comparisons of dentists and public health personnel.

The higher validities in the physicians group are perhaps due to factors intrinsic in the work situation. Such factors may be supervisors' relatively greater opportunity to observe carefully the work of medical personnel, particularly interns or lower grade officers under close supervision, and to develop evaluation standards and rating experience since physicians constitute the largest professional group in the Public Health Service.

Another possible explanation of the

validity coefficients obtained for the physicians group is that, in item analysis, a higher critical ratio was used as the standard for scoring in this group than in the other occupational groups. However, a recent study would seem to indicate that a stringent requirement for the level of discrimination of individual items does not necessarily increase total validity (2).

*Criteria.* Inspection of the validity coefficients (Table 3) from one criterion to another within the same level of supervisor shows that on all but the Forced Choice section higher validities occurred on Work Performance than on Personality. On the Forced Choice section, as high or higher validities were obtained on the Personality as on the Work Performance criterion for both levels of reporting supervisor, and for all occupational groups except the dentists, and the nurses rated by physicians.

A possible explanation of the observed differences in validity by criteria may be that a supervisor, if able to ascertain what constitutes a "good" and a "poor" rating (as is the case for rating scales and check lists), can more objectively evaluate the ratee on observable work-performance characteristics than he can on personal characteristics. When good and poor ratings are not so readily discernible, as presumably is the case with the forced choice type of evaluation, the objectivity of evaluations is perhaps increased so that they are as valid measures of the factors involved in a Personality as in a Work Performance criterion.

*Level of supervisor.* The validity coefficients in Table 3 may also be compared by level of supervisor. Since the primary Reports were used in item analysis, it is to be expected that they would yield higher validities than the secondary Reports. The median correlations in the table, however, indicate that validity held up surprisingly well on secondary Reports; the median correlations on the Personality criterion were even somewhat higher on the secondary than on the primary Reports. Considering the 42 pairs of coefficients which can be compared from one level of supervisor to another, primary Reports produced the higher validity in 21 comparisons, and secondary Reports produced the higher validity in the same number of comparisons. The median difference in coefficients was .07 for those comparisons in which primary Reports produced higher correlations, and .08 for those comparisons in which secondary Reports gave higher validities.

Differences in validity by level of supervisor, however, were observable within specific occupational groups. In 75 per cent or more of the comparisons of coefficients within the public health and the nurses-rated-by-nurses groups, the higher validity occurred on primary Reports. Since the primary Reports were used in item analysis, this finding is in the expected direction but in addition possibly reflects the fact that the primary, more immediate, supervisors of these groups are more likely to be in the ratees' profession than are the secondary supervisors. In 75 per cent or more of the comparisons within the medical and dental groups, the higher coefficients occurred on secondary Reports. It should be mentioned that hospitals and outpatient clinics are administered by a medical officer, and dental services are headed by a dental officer. For this reason, both the secondary and the primary supervisors of physicians and dentists are likely to be in either the same profession as the ratees or the same as the raters who performed criterion ratings. Further, secondary super-

visors of physicians and dentists are likely to be officers who routinely review efficiency reports and, therefore, have more information available as a basis for evaluating a ratee than do the primary supervisors.

In general, then, Reports completed independently by a second group of supervisors produced validities that compared favorably with those based on the Reports used in item analysis. It is likely that, had the secondary supervisors been at the same administrative level as the primary, differences in validity by level of supervisor which were apparent in specific occupational groups would have tended to occur less often; that is, the validities would be more nearly the same in all groups than occurred in the present data.

*Ratee grade.* A control on ratee grade was used both in the administration of criterion rating forms and in the establishment of occupational groups for purposes of item analysis. Nonetheless, since it was not feasible to control grade more precisely, this factor may be operating to increase spuriously the correlations shown in Table 3. In order to check this possibility, validity coefficients by grade were computed; small numbers of ratees made it necessary in some instances, however, to combine adjacent grades such as the senior and the director. Correlations by grade are presented in Table 6.

If grade were a systematic factor affecting both the criterion and Experimental Report variables, it would be expected that the validities based on all grades would be higher than the individual grade validities. That this is not the case may be seen from Table 6. There does not appear to be any consistent trend in the correlations as a function of grade level. The effect of combining grades was the masking of the higher validity obtained for some specific grades. In only one instance, in the nurses-rated-by-physicians group, was the correlation for all grades higher than any of the validity coefficients for the individual grades.

The significance of the difference in validity coefficients from one grade to another was tested. The relatively few comparisons, 16 out of a possible 122 (13.1 per cent), which yielded differences significant at the .05 level or below are shown in Table 6. No differences occurred in the public health group, and only two were found in the group of nurses rated by physicians.

In the medical group, the significant differences tended to involve higher validities in the Assistant grade as compared with other grades. Out of the 14 significant differences found in this professional group, 10 occurred in comparisons in which the higher coefficient was in the Assistant grade. This finding may be due to the fact that the majority of Assistant grade physicians are interns under close supervision; the supervisors of interns are experienced raters who have ample opportunity to observe the interns' performance.

The remaining four significant differences in the medical group occurred on the Reports completed by the secondary supervisors, Work Performance criterion, in the comparison of the combined Senior and Director grade with the Senior Assistant grade. The unusually high validities in the combined Senior and Director grade may have been due to the small number of ratees (about half the number available on primary Reports) and perhaps to a selective factor that resulted in the designation of highly experienced raters as the secondary supervisors for this grade.

Although differences in the validity coefficients found for the various ratee grades did occur, they were presumably the result of certain identifiable influences. The grade factor, as such, does not appear to have been operating in any systematic manner which spuriously increased the validities based on all grades.

### *Reliability of the Experimental Report*

*Rater agreement.* Correlations between scores from primary and secondary Reports, shown in Table 7, provide one

## TABLE 6

### Validity Coefficients for the Experimental Efficiency Report Based on Separate Grades

| Occupational Group | Grade | Primary Reports | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Work Performance Criterion | | | | | Personality Criterion | | | | |
| | | *N* | FC | JP | PQ | CL | *N* | FC | JP | PQ | CL |
| Nurses rated by physicians | Assistant | 48 | .64[c] | .45 | .04 | | 48 | .39 | .20 | .32 | |
| | Senior assistant | 28 | .23 | .02 | .10 | | 28 | .39 | .20 | .52 | |
| | All grades[a] | 92 | .44 | .28 | .43 | | 91 | .37 | .20 | .36 | |
| Physicians | Assistant | 83 | .64 | .71[c] | .65 | .66[c] | 83 | .62 | .55[c] | .59 | .57[b] |
| | Senior assistant | 108 | .54 | .53 | .50 | .45 | 108 | .65 | .35 | .42 | .46 |
| | Full | 58 | .59 | .45 | .47 | .45 | 58 | .63 | .34 | .54 | .45 |
| | Senior & Director | 70 | .61 | .65 | .57 | .48 | 71 | .50 | .24 | .42 | .20 |
| | All grades | 319 | .61 | .62 | .58 | .54 | 320 | .62 | .40 | .50 | .46 |
| Public health personnel | Full | 55 | .68 | .63 | .55 | .66 | 59 | .59 | .39 | .30 | |
| | Senior | 63 | .61 | .43 | .45 | .52 | 73 | .61 | .35 | .38 | |
| | All grades | 178 | .58 | .49 | .49 | .54 | 199 | .57 | .29 | .36 | |
| | | Secondary Reports | | | | | | | | | |
| Nurses rated by physicians | Assistant | 35 | .63 | .13 | .39 | | 35 | .51 | −.20 | .18 | |
| | Senior assistant | 24 | .53 | .49 | .22 | | 24 | .48 | .58[b] | .30 | |
| | All grades | 73 | .52 | .22 | .34 | | 72 | .45 | .10 | .27 | |
| Physicians | Assistant | 62 | .69 | .70[c] | .68[c] | .70[c] | 62 | .66 | .69[c] | .68 | .64 |
| | Senior assistant | 100 | .58 | .44 | .42 | .44 | 100 | .67 | .44 | .54 | .50 |
| | Full | 50 | .73 | .52 | .62 | .58 | 50 | .72 | .43 | .49 | .48 |
| | Senior & Director | 31 | .80[c] | .73[c] | .75[c] | .80[b] | 31 | .68 | .51 | .62 | .74 |
| | All grades | 243 | .67 | .58 | .60 | .60 | 243 | .68 | .52 | .58 | .56 |
| Public health personnel | Full | 46 | .46 | .56 | .50 | .51 | 49 | .53 | .48 | .56 | |
| | Senior | 42 | .44 | .34 | .44 | .55 | 47 | .65 | .24 | .39 | |
| | All grades | 127 | .37 | .37 | .37 | .40 | 139 | .50 | .30 | .33 | |

[a] The *N* for All Grades is based on all available cases including those in grades too small for the separate grade analysis.
[b] *r* is significantly higher at the .01 level than the underlined *r* in same column within the same report within the same officer group.
[c] *r* is significantly higher at the .05 level than the underlined *r* in same column within the same report within the same officer group.

## TABLE 7

### Correlations Between Scores from Primary and Secondary Reports[a]

| Occupational Group | Work Performance Criterion | | | | | Personality Criterion | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | No. Ratees | FC | JP | PQ | CL | No. Ratees | FC | JP | PQ | CL |
| Physicians | 243 | .57 | .51 | .58 | .61 | 243 | .60 | .51 | .58 | .59 |
| Public health personnel | 127 | .57 | .37 | .42 | .47 | 139 | .63 | .46 | .41 | |
| Research personnel | 107 | .59 | .53 | .62 | .54 | 111 | .64 | .54 | .61 | |
| Nurses rated by nurses | 91 | .59 | .39 | .55 | | 91 | .57 | .39 | .55 | |
| Nurses rated by physicians | 73 | .50 | .27[b] | .48 | | 72 | .62 | .31 | .48 | |
| Dentists | 42 | .65 | .55 | .52 | .58 | 43 | .64 | .55 | .52 | .61 |
| Median *r* | | .58 | .45 | .54 | .56 | | .63 | .49 | .54 | .60 |

[a] All *r*s not marked are significantly different from zero at the .01 level or below.
[b] *r* is significantly different from zero at the .05 level.

TABLE 8

RELIABILITY COEFFICIENTS FOR THREE SECTIONS OF THE EXPERIMENTAL EFFICIENCY REPORT

| Occupational Group | Forced Choice | | | | | Personal Qualifications[a] | | Check List | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | No. Ratees | No. Scored Tetrads | No. Scored Alternatives[b] | $r_{\frac{1}{2}I/II}$ | $r_{II}$ | $r_{\frac{1}{2}I/II}$ | $r_{II}$ | No. Scored Alternatives | $r_{\frac{1}{2}I/II}$ | $r_{II}$ |
| Physicians | 319 | 20 | 28 | .83 | .91 | .89 | .94 | 12 | .73 | .85 |
| P. h. personnel | 178 | 18 | 30 | .83 | .91 | .83 | .91 | 7 | .67 | .80 |
| Res. personnel | 131 | 18 | 22 | .78 | .88 | .90 | .95 | 9 | .68 | .81 |
| Nurses rated by nurses | 111 | 21 | 29 | .78 | .88 | .92 | .96 | | | |
| Nurses rated by phys. | 92 | 19 | 23 | .64 | .78 | .94 | .97 | | | |
| Dentists | 60 | 20 | 28 | .83 | .91 | .90 | .95 | 12 | .86 | .92 |
| Median $r_{II}$ | | | | | .90 | | .95 | | | .83 |

[a] The number of rating scales was eight, with the possible total stanine score ranging from 8 to 72.
[b] The number of scored alternatives rather than tetrads was used as the basis for computing reliability coefficients.

kind of measure of the reliability of the Experimental Report.

Correlations between Reports completed by the two groups of supervisors ranged from .27 to .65; over half were .55 or higher. Only the lowest correlation, .27, failed to reach the .01 level of significance; it was significant at the .05 level. It should be noted that the Job Proficiency (JP) section which consisted of a single rating scale tended to produce the lowest correlations. As the median $r$s indicate, the Forced Choice (FC) section tended to yield the highest correlations, although for the Work Performance criterion these did not differ markedly from those produced by the Personal Qualifications (PQ) and Check List (CL) sections. For the Personality criterion, the Forced Choice section gave the highest correlations in all instances, although the two correlations available on the Check List were of comparable size.

Since the primary and the secondary supervisors were at different administrative levels, the correlation coefficients are lower than might be obtained between Reports completed by different supervisors at the same level or between Reports completed by the same supervisor on two different occasions. Considering the factors operating to lower the coefficients, the correlations between scores on primary and secondary Reports, as measures of rater agreement, are fairly high.

*Spearman-Brown estimates.* From primary Reports, scored by the key developed against the Work Performance criterion, correlations between the odd and the even alternatives in each of three rating sections were corrected for length by the Spearman-Brown formula. A previous paper has reported that for Forced Choice tetrads, Spearman-Brown estimates of reliability were fairly close approximations of empirical reliabilities (5). Spearman-Brown estimates based on the present data are shown in Table 8. Since the Job Proficiency section involved only a single rating scale, it was not possible to compute a split-half coefficient on this part of the Report.

Considering the length of the various sections of the Report, the reliability coefficients are in the high range. As can be seen from Table 8, the number of scored alternatives on the Forced Choice and Check List sections varied somewhat for the different occupational groups. The number of rating scales in the Personal Qualifications section was the same (eight) for all groups.

The highest Spearman-Brown esti-

TABLE 9

COMPARISON OF EMPIRICAL VALIDITY COEFFICIENTS WITH VALIDITIES PREDICTED ON THE BASIS OF AN INCREASE IN LENGTH OF SCORING KEY

| Occupational Group | Section | No. Scored Alternatives | $r_{11}$ | Empirical Validity | Estimated Validity[c] | Limit of Validity[d] |
|---|---|---|---|---|---|---|
| Physicians | FC | 28 | .91 | .61[b] | | .64 |
| | PQ | 8 | .94 | .58 | .59 | .60 |
| | CL | 12 | .85 | .54 | .56 | .59 |
| Public health personnel | FC | 30 | .91 | .58 | | .61 |
| | PQ | 8 | .91 | .49 | .51 | .51 |
| | CL | 7 | .80 | .54 | .59 | .60 |
| Research personnel | FC | 22 | .88 | .65[a] | | .69 |
| | PQ | 8 | .95 | .39 | .40 | .40 |
| | CL | 9 | .81 | .59[a] | .63 | .66 |
| Nurses rated by nurses | FC | 29 | .88 | .42 | | .45 |
| | PQ | 8 | .96 | .39 | .40 | .40 |
| Nurses rated by physicians | FC | 23 | .78 | .44 | | .50 |
| | PQ | 8 | .97 | .43 | .43 | .44 |
| Dentists | FC | 28 | .91 | .53 | | .56 |
| | PQ | 8 | .95 | .55 | .56 | .56 |
| | CL | 12 | .92 | .50 | .51 | .52 |

[a] $r$ is significantly higher at the .01 level than the underlined $r$ in the same column within the same officer group.

[b] $r$ is significantly higher at the .05 level than the underlined $r$ in the same column within the same officer group.

[c] Estimated validity based on the same number of scored alternatives as the Forced Choice section.

[d] Limit of validity if report section were made infinitely long.

mates of reliability occurred on the Personal Qualifications (PQ) scales; all $r_{11}$'s were .91 or higher. Coefficients on the Check List (CL) ranged from .80 to .92, with a median of .83. The Forced Choice (FC) section yielded satisfactory reliabilities in all officer groups except the nurses-rated-by-physicians ($r_{11} = .78$); all other coefficients on this section were .88 or .91, with a median of .90.

*Validity as Related to Length of Scoring Key*

Since the sections of the Experimental Report were not equated in length, the fact that the Forced Choice was the longest section may account for its higher validity. The effect of length of scoring key on validity was tested on the Experimental Report sections for which both validity and reliability data were available from the primary Reports scored by keys developed on the basis of the Work Performance criterion. The results of the tests are shown in Table 9 which presents: (*a*) empirical validities and reliabilities and the number of scored alternatives in each Report section, re-

peated from previous tables for ease of comparison; (*b*) estimated validities for the Personal Qualifications (PQ) and the Check List (CL) sections, based on an increase in length of these sections to that of the Forced Choice (FC); and (*c*) the maximum validity that theoretically could be obtained on each section if it were made infinitely long (1, p. 166).

From Table 9, it may be seen that the Forced Choice section produced the highest empirical validity in all but one occupational group, the dental. Theoretically increasing the length of the other Report sections to that of the Forced Choice resulted in one additional group, public health personnel, in which the validity (estimated) of a section other than the Forced Choice was the highest. It is likely, however, that had the Report sections been equated for length, the three significant differences in validity (Table 9, footnotes a and b) that were obtained from one Report section to another would still have occurred.

If each of the Report sections were made infinitely long, it is apparent from a comparison of obtained validities with the theoretical limits of validity in Table 9 that the Personal Qualifications (PQ) section could be expected to show the smallest increase in validity (no more than .02). On the Forced Choice (FC) section, however, the increase to be expected ranges from .03 to .06, and on the Check List (CL), from .02 to .07.

While length of the various Report sections appears to have affected the magnitude of the validity coefficients, the greater number of scored alternatives in the Forced Choice (FC) section does not appear primarily responsible for the generally higher validity of this section. Although the Check List (CL) was the section most affected by the small number of scored alternatives, the evidence on limit of validity seems to indicate the relatively higher validity of the Forced Choice section.

*Multiple Correlations*

In order to determine the combination of Experimental Efficiency Report sections which would, for each occupational group, maximally predict the separate criteria, multiple correlation coefficients based on primary Reports were computed by the Wherry-Doolittle method of test selection (4, Chap. XIV). The intercorrelations on which the multiple correlation work was based are shown in Table 10. As can be seen from Table 10, the intercorrelations among Report sections were quite high, particularly in the medical and dental groups. In all occupational groups, the highest intercorrelations tended to occur between the Job Proficiency (JP) and Personal Qualifications (PQ) sections, with correlations ranging from .66 to .85. The Forced Choice (FC) and the Check List (CL) sections were also highly related, with a range in correlations from .68 to .82.

The multiple correlational data are presented in Table 11, which shows the $R$s based on selected predictors, the $\overline{R}$s obtained by application of the Wherry shrinkage formula, the validity coefficient of the best predictor in each occupational group, and the beta weights of predictors in the order in which the predictors were selected.

As Table 11 indicates, all $\overline{R}$s were significantly different from zero at the .01 level or below. Within each occupational group, the increase in validity as a result of using a team of Report sections rather than a single predictor may be seen by comparing the $\overline{R}$ with the $r$ of the first selected predictor. However, since the Wherry-Doolittle method of test selection does not guarantee that the increment due to the selection of successive variables significantly increases validity, the null hypothesis was tested by use of the $F$ ratio (3, p. 55). Tests for the significance

TABLE 10

INTERCORRELATIONS AMONG SECTIONS OF THE PRIMARY REPORTS*

| Occupational Group | Work Performance Criterion | | | | | | | Personality Criterion | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | *N* | FC-JP | FC-PQ | FC-CL | JP-PQ | JP-CL | PQ-CL | *N* | FC-JP | FC-PQ | FC-CL | JP-PQ | JP-CL | PQ-CL |
| Physicians | 319 | .64 | .70 | .79 | .85 | .69 | .73 | 320 | .53 | .68 | .74 | .85 | .65 | .73 |
| P. h. personnel | 178 | .56 | .56 | .68 | .78 | .59 | .49 | 199 | .52 | .52 | | .77 | | |
| Res. personnel | 131 | .51 | .47 | .77 | .79 | .48 | .45 | 136 | .38 | .45 | | .80 | | |
| Nurses rated by nurses | 111 | .41 | .60 | | .68 | | | 111 | .34 | .56 | | .68 | | |
| Nurses rated by phys. | 92 | .48 | .66 | | .66 | | | 91 | .33 | .58 | | .66 | | |
| Dentists | 60 | .73 | .70 | .82 | .80 | .75 | .76 | 62 | .71 | .63 | .75 | .80 | .76 | .74 |

* All *r*s are significantly different from zero at the .01 level or below.

of the difference in the $R^2$ (or $r^2$) based on the first selected predictor and the $R^2$ based on all selected predictors showed that, on the Work Performance criterion, validity was significantly increased at the .05 level or below in all groups except the dental. On the Personality criterion, a significant increase at the .05 level or below occurred in only the largest group, the medical.

As was previously noted in the discussion of the validity coefficients in Table 3, the data in Table 11 indicate that the validities obtained against the Work Performance criterion were higher than those on Personality, and that the physicians group was the one in which prediction was best.

The relative effectiveness of the four Report sections as measures of performance within each occupational group is evident from the results of the test selection. The predictors in Table 11 are those which, as determined by the Wherry-Doolittle method, combine to produce maximum multiple correlations. Inspection of the order in which predictors were selected and of the number of occupational groups in which they were selected shows that, considering both criteria, the Forced Choice (FC) section was a selected predictor in eleven of the twelve groups. Further, the Forced Choice section was the first selected predictor in nine of the groups. The Personal Qualifications (PQ) section was the next most frequently selected Report section occurring in eight of the officer groups, while the Job Proficiency (JP) section and the Check List (CL) each occurred in three groups. In general, the Forced Choice section in combination with one of the rating scale sections, usually Personal Qualifications, tended to produce the maximum correlations with the criteria.

TABLE 11

MULTIPLE CORRELATIONS BASED ON PRIMARY REPORTS

| Occupational Group | Work Performance Criterion | | | | | | Personality Criterion | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | No. Ratees | Selected Predictors | Beta Weights | $r$ of First Predictor | $R$ | $\overline{R}$[a] | No. Ratees | Selected Predictors | Beta Weights | $r$ of First Predictor | $R$ | $\overline{R}$[a] |
| Physicians | 319 | JP<br>FC | .39<br>.36 | .621 | .680[c] | .679 | 320 | FC<br>PQ<br>CL | .57<br>.19<br>−.10 | .624 | .636[b] | .633 |
| Public health personnel | 178 | FC<br>PQ<br>CL | .31<br>.21<br>.23 | .580 | .637[c] | .632 | 199 | FC<br>PQ | .52<br>.09 | .571 | .576 | .573 |
| Research personnel | 131 | FC<br>CL<br>JP | .43<br>.20<br>.13 | .650 | .674[b] | .668 | 136 | FC | .63 | .631 | .631 | .631 |
| Nurses rated by nurses | 111 | FC<br>PQ | .29<br>.22 | .420 | .455[b] | .447 | 111 | FC<br>PQ | .34<br>.16 | .430 | .451 | .442 |
| Nurses rated by phys. | 92 | FC<br>PQ | .28<br>.25 | .442 | .482[b] | .472 | 91 | FC<br>PQ | .23<br>.23 | .367 | .411 | .399 |
| Dentists | 60 | PQ<br>FC | .34<br>.30 | .545 | .584 | .575 | 62 | JP | .41 | .413 | .413 | .413 |

[a] All $\overline{R}$s are significantly different from zero at the .01 level or below.
[b] $R$ is significantly higher than $r$ at the .05 level.
[c] $R$ is significantly higher than $r$ at the .01 level or below.

Since the multiple correlational work was based on validity coefficients obtained for item-analysis samples, the findings concerning the relative effectiveness of the various Report sections and the sizes of the multiple correlation coefficients require verification on independent samples. Evidence from the "cross" scoring keys, however, indicated that little decrease is to be expected in cross validation of the Forced Choice section of the Report. Further, cross validation of the Personal Qualifications and Job Proficiency scales was not deemed necessary since matched samples had produced highly similar scoring keys. In view of these considerations, it is likely that the validity coefficients will not show a marked decrease in subsequent samples.

*Validity of the Officer's Progress Report*

Validity coefficients based on the Rating Scales (RS) and the Narrative Comments (NC) sections of the Officer's Progress Report and on sections of the Experimental Report completed by the primary supervisors are shown in Table 12. Considerable attrition in the number of Experimental Reports occurred as a result of using only those ratees on whom both Reports were available. The median correlations for the Experimental Report shown in Table 12, however, are about the same size as the corresponding median validities in Table 3.

Tests of the significance of the difference in the validity coefficients in Table 12 from one Report section to another revealed that coefficients in 20 per cent (30 out of a possible 150) of the comparisons differed significantly at the .05 level or below. The specific comparisons which produced significant differences are shown in Table 13. The percentage of significant comparisons was less than occurred in the tests of differences on the Experimental Report (see Table 4); this was probably due to the smaller number of cases available on the combined Reports.

As was previously found, the Forced Choice

TABLE 12

VALIDITY COEFFICIENTS FOR THE EXPERIMENTAL REPORT AND THE OFFICER'S PROGRESS REPORT[b]

| Occupational Group | Work Performance Criterion | | | | | | | Personality Criterion | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $N$[a] | Experimental Report (Primary Reports) | | | | Progress Report | | $N$[a] | Experimental Report (Primary Reports) | | | | Progress Report | |
| | | FC | JP | PQ | CL | RS | NC | | FC | JP | PQ | CL | RS | NC |
| Physicians | 187 | .62 | .69 | .61 | .54 | .55 | .57 | 187 | .59 | .52 | .56 | .47 | .52 | .53 |
| P. h personnel | 108 | .64 | .49 | .47 | .50 | .56 | .52 | 120 | .59 | .23[c] | .30 | | .32 | .33 |
| Res. personnel | 74 | .74 | .56 | .49 | .65 | .58 | .69 | 73 | .72 | .15[d] | .30 | | .35 | .33 |
| Nurses rated by nurses | 88 | .43 | .34 | .50 | | .40 | .20[d] | 88 | .46 | .24[c] | .38 | | .34 | .17[d] |
| Nurses rated by phys. | 72 | .49 | .34 | .52 | | .57 | .26[c] | 71 | .38 | .19[d] | .39 | | .27[c] | .15[d] |
| Dentists | 40 | .52 | .47 | .57 | .50 | .48 | .46 | 42 | .42 | .47 | .44 | .36[c] | .41 | .44 |
| Median $r$ | | .57 | .48 | .51 | .55 | .56 | .49 | | .53 | .24 | .39 | .42 | .35 | .33 |

[a] Based on the number of officers on whom both Experimental and Progress Reports were available.
[b] All $r$s not marked are significantly different from zero at the .01 level or below.
[c] $r$ is significantly different from zero at the .05 level. [d] $r$ does not reach the .05 level of significance.

TABLE 13

COMPARISONS OF SECTIONS OF THE EXPERIMENTAL REPORT AND THE OFFICER'S PROGRESS REPORT IN WHICH VALIDITY COEFFICIENTS DIFFERED SIGNIFICANTLY

| Occupational Group | Work Performance Criterion | Personality Criterion |
|---|---|---|
| | Report Sections Compared[a] | Report Sections Compared[a] |
| Physicians | FC vs. CL[c]<br>JP vs. PQ[b]<br>JP vs. CL[b]<br>JP vs. RS[b]<br>JP vs. NC[c] | FC vs. CL[b]<br>PQ vs. CL[c] |
| Public health personnel | FC vs. JP[c]<br>FC vs. PQ[c] | FC vs. JP[b]<br>FC vs. PQ[b]<br>FC vs. RS[b]<br>FC vs. NC[b] |
| Research personnel | FC vs. JP[c]<br>FC vs. PQ[b]<br>FC vs. RS[c]<br>NC vs. PQ[c] | FC vs. JP[b]<br>FC vs. PQ[b]<br>FC vs. RS[b]<br>FC vs. NC[b]<br>PQ vs. JP[c] |
| Nurses rated by nurses | FC vs. NC[c]<br>PQ vs. JP[c]<br>PQ vs. NC[b] | FC vs. NC[c] |
| Nurses rated by physicians | FC vs. NC[c]<br>RS vs. JP[c]<br>PQ vs. NC[c]<br>RS vs. NC[b] | |

[a] The Report section on which the higher validity occurred is listed first.

[b] Validity coefficients for the sections compared differ significantly at the .01 level or below.

[c] Validity coefficients for the sections compared differ significantly at the .05 level.

(FC) section produced more (18 out of 30) of the significantly higher validity coefficients than any other Report section. In no instance was a validity coefficient on the Forced Choice section significantly lower than that of another section. The number of comparisons in which each of the remaining five Report sections produced a validity significantly higher than another section ranged from none on the Check List to five on the Personal Qualifications section.

From the results of the significance of difference tests and from the median validity coefficients, it is interesting to note that the two sections of the Progress Report, Rating Scales (RS) and Narrative Comments (NC), produced validities that compare favorably with all sections of the Experimental Report except the Forced Choice.

The data on the Officer's Progress Report again suggest the relative superiority of the forced choice type of evaluation as compared with more conventional rating methods. However, since the Progress Report was completed under operational rather than experimental conditions, no attempt will be made to compare the two Reports by use of multiple correlational techniques. It is anticipated that in a later study, it will be possible to collect data on the Progress Report along with cross-validational data for the Experimental Report so that a more intensive comparison of the two Reports can be made.

## SUMMARY

This study has compared the relative efficacy of the forced choice technique with other more conventional evaluation methods as measures of the performance of professional health personnel working as commissioned officers in the United States Public Health Service.

Four sections of an Experimental Efficiency Report were studied: (*a*) 50 Forced Choice tetrads adapted from items developed by the Department of the Army; (*b*) a ten-point scale for rating a ratee's Job Proficiency in his primary job function; (*c*) eight ten-point scales for the evaluation of Personal Qualifications; and (*d*) a twenty-two-item Check List developed from comments appearing in the Officer's Progress Report, the efficiency report in operational use in the Service. In addition, two sections from the Officer's Progress Report were available for comparison with those in the Experimental Report: (*a*) eleven five-point Rating Scales for evaluating

various aspects of performance in the Public Health Service; and (*b*) Narrative Comments coded and scored by a method previously developed.

The criteria of Service performance were twenty-point graphic rating scales for the evaluation of Work Performance and Personality. A ratee's criterion score was the average of the ratings given him by his work associates on each criterion. The results of the study have shown that:

1. The Forced Choice section of the Experimental Report was highly effective for evaluating the performance of professional personnel commissioned in the Public Health Service.

Of 24 validity coefficients based on scoring keys developed by selecting the best tetrads from those which had the same empirically determined scoring weights in independent matched samples, 41.7 per cent were .62 or higher. All except one of the coefficients were significant at the .01 level or below; this one was significant at the .05 level (Table 2, "combined scoring").

Only 12.5 per cent of 24 validity coefficients based on item-analysis samples showed a significant decrease at the .05 level or below in cross validation (Table 2, comparison of "self" and "cross" scoring).

2. The validity of the Forced Choice section was generally higher than that of the other Report sections studied.

Out of 36 significant differences (.05 level or below) obtained in comparisons of the validity of the Experimental Report sections, 27 (75 per cent) involved higher validities on the Forced Choice tetrads, while only one involved a lower coefficient on this section (Table 4).

The Forced Choice section contained a greater number of scored alternatives than the other sections of the Experimental Report; estimates of validity based on theoretically making each section infinitely long, however, seemed to indicate that the length of the Forced Choice section was not primarily responsible for its generally higher validity (Table 9).

Out of 12 multiple correlation coefficients computed on the Experimental Report by the Wherry-Doolittle method of test selection, 11 involved the Forced Choice section as a selected predictor; in nine of the 11, this section was the first selected predictor (Table 11).

Comparisons of the validities of sections of both the Experimental Report and the Officer's Progress Report revealed 30 significant differences; 18 (60 per cent) involved higher validities on the Forced Choice tetrads while none involved a lower coefficient on this section (Table 13).

3. Of six occupational groups for which separate scoring keys were developed for the Experimental Report, the largest group, that of hospital physicians, was the one in which the highest validity coefficients generally occurred. The occupational groups, other than physicians, which were involved in the study were dentists, public health personnel, research personnel, and nurses rated by two different criterion rater groups, physicians and nurses.

Of 44 significant differences (.05 level or below) obtained in comparisons of validity coefficients from one occupational group to another on sections of the Experimental Report, 33 (75 per cent) involved higher coefficients in the medical group (Table 5).

Multiple correlations ($\overline{R}$) for the Experimental Report computed by the Wherry-Doolittle method were, for the medical group, .68 and .63 against the Work Performance and the Personality criteria, respectively. Both coefficients were significant at the .01 level or below, and both represented a significant increase (.05 level or below) in validity over that obtained on the second-best single Report section. Multiple correlations in the public health and the research groups were also relatively high, ranging from .57 to .67 on the two criteria (Table 11).

4. Validity coefficients were generally higher when Work Performance rather than Personality was used as the criterion.

On all sections of the Experimental Report except the Forced Choice, higher validities were obtained with the Work Performance criterion than with the Personality criterion. Forced Choice validities were not consistently higher for either criterion (Table 3).

Within each officer group, a higher multiple correlation coefficient was obtained for the Experimental Report when Work Performance was used as the criterion than when Personality was used (Table 11).

Considering sections from both the Experimental and the Progress Reports, higher

validities occurred in all but one instance when Work Performance rather than Personality was used as the criterion (Table 12).

5. Experimental Reports completed by a group of supervisors independent of and at a higher administrative level than those completing the Reports used in item analysis produced validities that compared favorably with the validities of the item-analysis Reports.

Of 42 comparisons of validity from one level of supervisor to another, 21 involved higher validity coefficients on item-analysis Reports, and 21 involved higher validities on Reports completed by an independent group of supervisors. The median difference in validity coefficients in those comparisons in which item-analysis Reports yielded the higher validities was .07, and in those in which the independent Reports gave higher coefficients, .08 (Table 3).

6. Validity coefficients based on sections of the Experimental Report did not show a consistent trend as a function of grade level.

Of 122 possible comparisons of validity coefficients from one grade to another, only 16 (13.1 per cent) yielded differences significant at the .05 level or below. Validity coefficients for separate grades were also compared with those based on all grades. The effect of combining grades appeared to be the masking of the higher validity obtained in certain specific grades; in only one instance was a combined grade validity higher than any of the coefficients for the separate grades (Table 6).

7. The sections of the Experimental Efficiency Report exhibited satisfactory reliabilities.

Spearman-Brown estimates of reliability for three of the Report sections ranged from .78 to .97 Median reliabilities ($r_{11}$) were .95, .90, and .83, respectively, for the Personal Qualifications, the Forced Choice, and the Check List sections. It was not possible to compute a split-half coefficient for the Job Proficiency section since it consisted of a single rating scale (Table 8).

As a measure of rater agreement, scores on Reports completed by two groups of supervisors at different administrative levels were correlated. Over half of the correlations between the two sets of Reports were .55 or higher (Table 7).

8. The Rating Scales and Narrative Comments sections of the Officer's Progress Report appeared to be about as adequate measures of performance as sections of the Experimental Report other than the Forced Choice.

Median validity coefficients for the Progress Report compared favorably with those for sections of the Experimental Report other than the Forced Choice. Data on the two Reports, however, were collected under different conditions so that comparative results are viewed as tentative (Table 12).

The significance of the difference was tested in the validity coefficients obtained for the various sections of both Reports. Significantly higher (.05 level or below) validities occurred on each section of the Progress Report about as frequently as on the Experimental Report sections other than the Forced Choice (Table 13).

9. Multiple correlations computed on the Experimental Report indicated that prediction was in some instances, but not in others, increased by the use of more than one Report section.

All multiple correlations were significantly different from zero at the .01 level or below. Five of the six correlations based on the Work Performance criterion represented a significant increase (.05 level or below) in validity over that obtained on the best single Report section for each officer group. Only one of those based on the Personality criterion, however, showed such a significant increase (Table 11).

10. The combination of sections of the Experimental Report which produced the maximum correlation with the criteria, as determined by the Wherry-Doolittle method, differed for each of the officer groups studied, but tended to include the Forced Choice in combination with one of the rating scale sections, usually Personal Qualifications.

Of 12 multiple correlations computed, six involved the Forced Choice and Personal Qualifications sections as the only selected predictors, and three involved these two sections in combination with a third section. In one multiple correlation the Forced Choice section was selected in combination with the Job Proficiency scale, and each of these sections was the only predictor selected in the two remaining multiple correlations (Table 11).

## Implications of the Findings

Evaluation of the performance of highly trained professional personnel poses a difficult measurement problem. The complexities of the work requirements for such personnel make adequate, objective criteria of professional competency difficult to obtain at the present time. Any criterion or criteria should presumably reflect such personal characteristics as professional knowledge, judgment, technical skill, originality, emotional adjustment, and ability to administer programs in a professional specialty. While the inadequacies of the type of criterion employed in the present work are recognized, practical considerations necessitated the use of a conventional work-associates' rating method.

With the type of item analysis and control of experimental variables used in this study, it would appear that, within the limitations imposed by a rating criterion, satisfactory validity and reliability of performance evaluation methods for professional health personnel can be obtained. Of particular interest are the results obtained for the forced choice tetrads which, under the conditions of this study, generally produced higher validity coefficients than other methods of assessing or reporting efficiency. Since rating-scale methods of efficiency reporting have widespread usage, it may also be of general interest that these methods produced satisfactory validity as measures of professional performance.

The findings appear to be applicable to other organizations employing medical, scientific, and other health personnel similar to those employed by the Public Health Service. With regard to the forced choice items, it may be recalled that the items used here, although scored by keys developed from item analysis of Experimental Efficiency Reports completed on Public Health Service personnel, were developed in another organization on an employee population quite different from that of the Public Health Service. From the evidence concerning validity of the tetrads in the variety of work activities in the Public Health Service (medical care, research, and public health), it may be inferred that the item content and the technique are such as to be relevant in a number of different employment situations.

## APPENDIX

## A. SAMPLES OF ITEMS FROM SECTIONS OF THE EXPERIMENTAL EFFICIENCY REPORT

### *Section I. Forced Choice*

**Directions for completing: From each of the following sets of four words or phrases, mark the one word or phrase in each set which is "most descriptive" and the one which is "least descriptive" of the officer you are rating.**

| | Most | Least |
|---|---|---|
| A. A go-getter who always does a good job | | |
| B. Cool under all circumstances | | |
| C. Doesn't listen to suggestions | | |
| D. Drives instead of leads | | |

| | Most | Least |
|---|---|---|
| A. Cannot assume responsibility | | |
| B. Knows how and when to delegate authority | | |
| C. Offers suggestions | | |
| D. Too easily changes his ideas | | |

| | Most | Least |
|---|---|---|
| A. Modest and reserved | | |
| B. Doesn't have the drive or force he should | | |
| C. Antisocial | | |
| D. Respected by all fellow officers | | |

### *Section II. Job Proficiency*

**Directions for completing: From the Service functions listed below, select the one you consider to be the primary job of the officer you are evaluating. Rate the officer's job proficiency in this function by marking a position on the ten-point scale.**

1. Operation in a technical or specialized Public Health program
2. Care of patients or furnishing services to patients
3. Administration of a clinical or medical care program at any level
4. Directly performing research work

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| FOR RATING OFFICER<br>☐ Number of Function | | | | | | | | | | |

### *Section III. Personal Qualifications*

**Directions for completing: By marking a position on a ten-point scale, rate the officer on each of the following personal qualifications.**

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| The degree to which he is able to discriminate & evaluate facts to arrive at logical conclusions. | | | | | | | | | | |
| The degree to which his appearance and behavior cause people to react favorably. | | | | | | | | | | |
| The degree to which he is able to carry out orders with consistency & firmness to achieve objectives | | | | | | | | | | |

### *Section IV. Check List*

Directions for completing: From the following statements, determine whether or not each statement applies to the officer under consideration. If a statement does apply, mark space one (1); if it does not, mark space two (2).

| Applies | Does not Apply | |
|---|---|---|
| 1 = | 2 = | This officer has a broad and detailed knowledge of his profession |
| 1 = | 2 = | This officer's usefulness is limited to a narrow field |
| 1 = | 2 = | This officer does an excellent job of planning and organizing his work |

## B. SAMPLES OF ITEMS FROM SECTIONS OF THE OFFICER'S PROGRESS REPORT

### *Rating Scales*

| Indicate rating by check mark | Unsatisfactory | Fair | Good | Very Good | Excellent |
|---|---|---|---|---|---|
| Judgment | ____ | ____ | ____ | ____ | ____ |
| General professional knowledge | ____ | ____ | ____ | ____ | ____ |
| Proficiency in assigned duties | ____ | ____ | ____ | ____ | ____ |
| Tact | ____ | ____ | ____ | ____ | ____ |
| General fitness for the service | ____ | ____ | ____ | ____ | ____ |

### *Questions Eliciting Narrative Comments*

Are you satisfied to have this officer? Yes ☐ No ☐ Give reasons .....

Handicaps .....

What are your recommendations for this officer's improvement? .....

Remarks .....

## REFERENCES


1. ADKINS, DOROTHY C. *Construction and analysis of achievement tests.* Washington, D.C.: U. S. Government Printing Office, 1947.
2. APPEL, V., & KIPNIS, D. The use of levels of confidence in item analysis. *J. appl. Psychol.*, 1954, **38**, 256–259.
3. DWYER, P. S. The relative efficacy and economy of various test selection methods. *Psychol. Res. Section Rep.* No. 957, Adjutant General's Office, 1952.
4. GARRETT, H. E. *Statistics in psychology and education.* New York: Longmans, Green, 1947.
5. HARRIS, F. J., HOWELL, M. A., & NEWMAN, S. H. Forced choice tetrads—effect of scoring procedure and key length on validity and reliability. *Educ. & psychol. Measmt*, **16**, 454–464.
6. MCNEMAR, Q. *Psychological statistics.* New York: Wiley, 1949.
7. NEWMAN, S. H. The officer selection and evaluation program of the U. S. Public Health Service. *Amer. J. publ. Hlth*, 1951, **41**, 1395–1402.
8. NEWMAN, S. H. Quantitative analysis of verbal evaluations. *J. appl. Psychol.*, 1954, **38**, 293–296.
9. NEWMAN, S. H., BUSSEY, R., & EPSTEIN, M. Performance criteria and evaluation methods for professional health personnel. Unpublished manuscript, 1955.

10. SISSON, D. E. Forced-choice—The new army rating. *Personn. Psychol.*, 1948, 1, 365–381.
11. THURSTONE, L. L. Attitudes can be measured. *Amer. J. Sociol.*, 1928, 33, 529–554.
12. WITSELL, E. J. The new officer efficiency report. *The Reserve Officer*, 1947, 24, 8–10.
13. Major study of comparative validity of five periodic officer efficiency reporting methods. *Personn. Res. Section Rep.* No. 670, Adjutant General's Office, 1945.
14. Studies of Officer Efficiency Report, WD AGO Form 67-1, in operation. I. Revalidation. *Personn. Res. Section Rep.* No. 791, Adjutant General's Office, 1949.

*(Accepted for publication February 17, 1957)*